

# Weekly Report

July 9, 2017

## 1 Work

本周我对于目前对tsne降维方法加速的方法做了一些简单的调研，LargeVis从算法复杂度上来说已经达到 $O(N)$ 的地步，因为他采用了负采样算法（word2vec中使用的）。然而看了近几年的文章，也没有对负采样算法有速度上的改进，只能通过分布式计算等方法进行加速，都没有对复杂度进行改进。本周我基本实现了LargeVis的GPU加速，目前大概需要90秒，相比于原来的算法，在优化上面目前大约有2-3倍的加速。

## 2 Paper Reading

### 2.1 The State of the Art in Integrating Machine Learning into Visual Analytics

本文主要对于机器学习使用在可视分析领域的文献归类。在可视分析中机器学习方法主要分为四种：降维，聚类，分类和回归。可视分析和机器学习结合的部分主要是在交互上，可以做1) 修改参数和缩小计算空间；2) 定义用户想要的期望结果。

### 2.2 The State-of-the-Art in Predictive Visual Analytics

与上一篇文章类似，本文也是讲机器学习和可视化的结合，主要是从机器学习的流程中可视分析可以介入的过程：1) 数据预处理；2) 特征抽取；3) 建模；4) 结果探索；5) 效果验证。

## 2.3 Distributed Negative Sampling for Word Embeddings

这篇文章实现的是一个在分布式集群中完成词嵌入的负采样优化算法。主要的贡献是将原来需要在整个数据集上运行的负采样算法（因为每一次的迭代最好能够在最新的数据位置上进行）分解为可以在单个节点上进行的算法，从而减少节点之间的交流通信，减少因为网络延迟而导致的计算停滞。

## 2.4 ivhd: A fast and simple algorithm for embedding large and high-dimensional data

MDS要保持的是在高维空间中的距离和在低维空间中的距离尽量保持一致，这和tsne其实比较类似（tsne需要保持的是概率）。然而计算所有点的之间的距离的代价是 $O(N^2)$ 。作者在这篇文章中使用k个最近邻和一个随机数据点作为算法要保持的距离，这样可以显著提高计算速度。虽然在某些数据集上面会产生一些问题，比如同一类不是聚拢在一起，但总体来说还是对MDS有了非常大的提速。

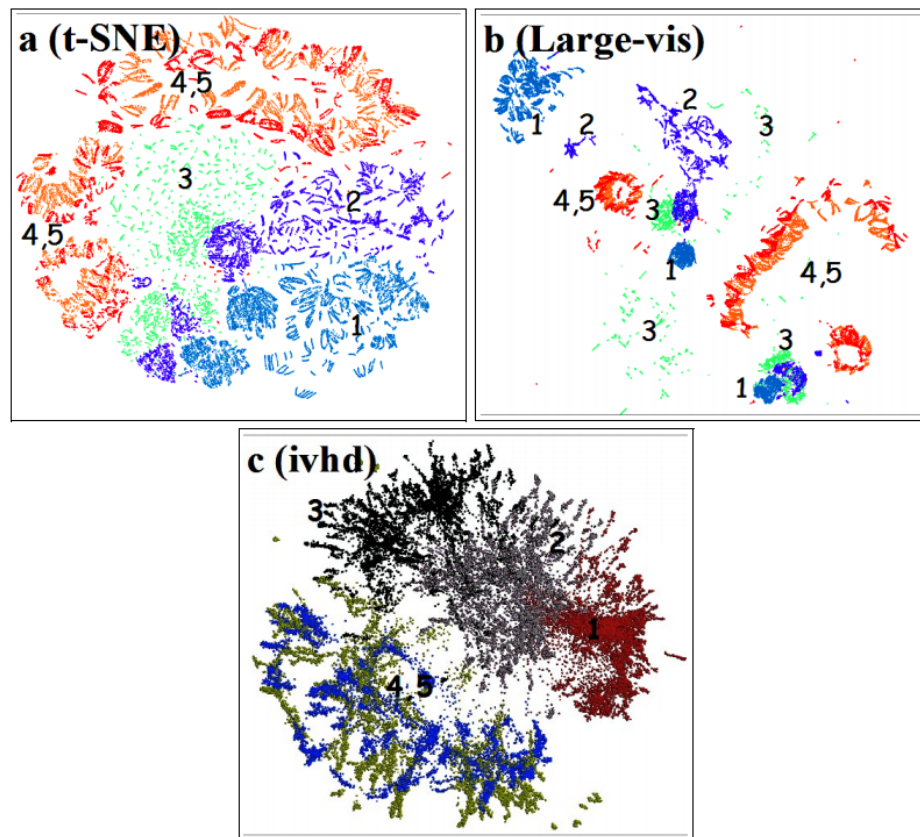


Figure 1: